Using a natural language and gesture interface for unmanned vehicles

Dennis Perzanowski*[a], Alan C. Schultz[b], William Adams[b], and Elaine Marsh[a]

Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
[a]Code 5512 and [b]Code 5514
Washington, DC 20375-5337

ABSTRACT

Unmanned vehicles, such as mobile robots, must exhibit adjustable autonomy. They must be able to be self-sufficient when the situation warrants; however, as they interact with each other and with humans, they must exhibit an ability to dynamically adjust their independence or dependence as co-operative agents attempting to achieve some goal. This is what we mean by adjustable autonomy. We have been investigating various modes of communication that enhance a robot's capability to work interactively with other robots and with humans. Specifically, we have been investigating how natural language and gesture can provide a user-friendly interface to mobile robots. We have extended this initial work to include semantic and pragmatic procedures that allow humans and robots to act co-operatively, based on whether or not goals have been achieved by the various agents in the interaction. By processing commands that are either spoken or initiated by clicking buttons on a Personal Digital Assistant and by gesturing either naturally or symbolically, we are tracking the various goals in the interaction, the agent involved in the interaction, and whether or not the goal has been achieved. The various agents involved in achieving the goals are each aware of their own and others' goals and what goals have been stated or accomplished so that eventually any member of the group, be it a robot or a human, if necessary, can interact with the other members to achieve the stated goals of a mission.

Keywords: adjustable autonomy, gesture, goal tracking, mixed-initiative, natural language interface, robotics

1. INTRODUCTION

To achieve any level of independence, autonomy, and/or cooperation between humans and robots in completing a task, the interaction should allow either humans or robots to be the originators of goals and motivations. We refer to systems that permit this type of interaction as *mixed-initiative* systems.

In the context of mixed-initiative systems, *adjustable autonomy* is a critical requirement. Systems exhibiting this feature permit participants to interact with dynamically varying levels of independence, intelligence, and control. In these systems, human users and robots interact freely and cooperatively to achieve their goals. There is no master/slave relationship. Participants may adjust their level of autonomy as required by the current situation. This requires that participants are aware of what the goal is and how each can contribute to achieve that goal effectively.

The above type of interaction has a parallel in natural language. Dialogs between people exhibit the same type of cooperative behavior to achieve certain goals and exchange of information. We have been utilizing a dialog-based[1.] approach to drive our research on achieving adjustable autonomy in a robotics domain.

Our research addresses the case of human-robot interactions that require close interaction, and it was for this reason that we incorporated a natural language interface. Our research on the natural language and gestural interface is based upon the premise that people communicate with other people easily and they use natural language and physical gesturing to do so, among other channels of communication, such as facial expression. We, therefore, assumed that people might readily interact with autonomous robots in much the same fashion; namely, by using speech and body gestures.

* Correspondence: Email: dennisp@aic.nrl.navy.mil; WWW: http//www.aic.nrl.navy.mil/~dennisp; Telephone: 202-767-9005; Fax: 202-767-3172

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
2000

2. REPORT TYPE

3. DATES COVERED
-

4. TITLE AND SUBTITLE
Using a natural language and gesture interface for unmanned vehicles

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Navy Center for Applied Research in Artificial Intelligence,Naval Research Laboratory,Code 5510,Washington,DC,20375

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
Unmanned vehicles, such as mobile robots, must exhibit adjustable autonomy. They must be able to be self-sufficient when the situation warrants; however, as they interact with each other and with humans, they must exhibit an ability to dynamically adjust their independence or dependence as co-operative agents attempting to achieve some goal. This is what we mean by adjustable autonomy. We have been investigating various modes of communication that enhance a robot’s capability to work interactively with other robots and with humans. Specifically, we have been investigating how natural language and gesture can provide a user-friendly interface to mobile robots. We have extended this initial work to include semantic and pragmatic procedures that allow humans and robots to act co-operatively, based on whether or not goals have been achieved by the various agents in the interaction. By processing commands that are either spoken or initiated by clicking buttons on a Personal Digital Assistant and by gesturing either naturally or symbolically, we are tracking the various goals in the interaction, the agent involved in the interaction, and whether or not the goal has been achieved. The various agents involved in achieving the goals are each aware of their own and others’ goals and what goals have been stated or accomplished so that eventually any member of the group, be it a robot or a human, if necessary, can interact with the other members to achieve the stated goals of a mission.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**7**	

Of course, the number of channels of communication that are available in human communication is much more numerous than those we have elected to concentrate on here. We have limited our considerations to basically two, spoken natural language and body gestures of two types which we outline below, in order to determine the empirical consequences of using these two channels of communication in a human-robot interface.[2.]

Since our initial research involved directing robots to certain locations, we immediately saw the need to incorporate gesture into our natural language interface. However, natural language contains inherent ambiguities when it comes to processing location and directional information. Thus, for example, without some sort of accompanying gesture to indicate the exact place in the real world where the speaker intends the hearer to move, a sentence like "Go over there" is ambiguous. With an accompanying hand/arm movement, a nod of the head, or even a sidelong glance in a particular direction, the utterance can be processed along with the needed locative information obtained from the gesture. We also saw the need to incorporate corrective actions when, for example, a user might say "Turn left" and point to his/her own left, which from the point of view of the hearer is an incorrect direction. Therefore, we incorporated error correcting procedures which were informative. Telling the user that an error was obtained is not informative. Telling the user "You told me to turn left but pointed in another direction." assists both participants in the dialog to achieve their desired goals. We further believe this is more in keeping with how humans might interact in such a situation, and it is certainly more informative than generating some sort of obscure error code.

During the initial development stage of the interface, we also found it necessary to incorporate fragmentary input. People do not always speak in what are grammatically known as "complete" sentences. Many of our verbal interactions can be classified as sentence fragments. Such as in (1-3).

(1) Go over there. (no accompanying gesture provided)
(2) Where?
(3) Over there? (accompanying gesture provided)

But this adds an additional load onto the natural language processing. For sentence (3) to be processed, the natural language components of the interface need to know what the missing material in the sentence is. In the case of (1-3), the missing material can be obtained from (1), namely the verb "go." The rules of discourse that permit this will not concern us here, but suffice it to say that such processing is not unique but is a linguistically motivated way of acquiring the necessary information in a dialog. Suffice it to say here, we periodically and temporarily store the predicate or verbal information of sentences in the dialog, as well as the corresponding arguments of each predicate. This permits the processing of the necessary information. Again, when and for how long such information is kept as a matter of record is motivated by principles of discourse, and basically, it hinges on whether or not a goal has been achieved or not. In other words, information from (1) is kept on hand for as long as it is needed. If (1) is acted upon, because all of the information is present and the sentence is unambiguous, then the information need no longer be kept or stored. Thus, in the example above, where no accompanying gesture was provided with (1), the system stores the predicate information (we will go into greater detail below), until such time as the goal is actually achieved, or changed as the case may be.

So, out of a consideration of processing incomplete verbal input, we discovered that we needed to track goals and whether they were achieved or not. We needed to keep information on hand, just in case a fragmentary input was obtained. But then in the interest of streamlining the interface, we needed to know how long this information should be kept. It basically boiled down to goal attainment. Once attained, information could be cleared, thereby providing us with additional storage space for subsequent interactions. But as an interesting side effect, keeping information on hand, provided us with a way of allowing our robotic system to be more of an independent and cooperative agent in exchanges. Having the information about which goal was achieved or not allowed our robotic system to interrupt activity for whatever reason, and then return to previous actions whenever they had not been obtained. Again, human-human interactions are characterized by interruptions, reversals and completions of tasks on hand, and tasks yet to be accomplished but previously requested. Therefore, we felt that incorporating this capability into our interface led to a more cooperative interaction between our human users and robotic system. Thus, as our interface development progressed, we saw a need to integrate the language and gestural capabilities of the interface by tracking goals in human/robot interactions.[3.] We argued that tracking goals provided us with a means of achieving varying levels of autonomy.

Recently, we included a mechanical means of communication with the robots via palm devices.[4.] Our reason for doing so was to expand the gesture capabilities of our interface. While we were initially interested in natural commands and gestures, we felt it necessary to be able to incorporate what we call "synthetic" means of interaction. Natural gestures we consider to be those made by natural movements of a person's arm and/or hand. Synthetic gestures are those made by pointing and clicking on a mechanical device, such as clicking on buttons on a Personal Digital Assistant, hereafter PDA, display and drawing or pointing to locations on a PDA map. We were motivated to incorporate these additional modalities because we felt that solely natural modes of interaction were somewhat limiting and constraining the human user.

In some situations, for example, people might not want to be heard or perhaps might not be able to be heard while communicating with an unmanned vehicle or robotic device. Also, people might not want their gestures to be visible to other individuals who might be watching the interaction. For example, soldiers and other personnel involved in urban guerilla warfare situations might not want their positions to be discerned by the enemy. Talking out loud and waving their hands would be literally a dead giveaway. We, therefore, incorporated modes by which the same information could be communicated between the interacting agents, so they had several ways now in which they could communicate. Choice of which mode to use to interact is entirely up to the participants, and modes can be mixed or matched. Thus, for example, a verbal command can be matched with a synthetic gesture on the map of the PDA. Likewise, individuals can issue a command by clicking on a button on a PDA and gesture naturally. We do not restrict the users in any way. The system takes the input by first determining the source, which is basically immaterial, and then processes the input in a similar way whether the input comes from either a natural or a synthetic source.

Thus, we have expanded the kinds of interactions permitted in our interface, but we further argue that these added capabilities require the various components to be tightly integrated in order to achieve success. Our research has brought us to the conclusion that using a dialog-based approach to communication as the guiding principle creates the level of integration necessary for such a system to function. Furthermore, the integration of multiple modes of communication using a dialog-based approach favorably affects adjustable autonomy. However, before considering the interface in greater detail, let us look at some additional self-imposed limitations on the system.

When people talk, they gesture. Some of those gestures are meaning-bearing, while others are superfluous, some redundant, and some indicate an emotional or intentional state of the speaker. We limit ourselves to the meaning-bearing gestures that disambiguate locative elements of spoken natural language or on a synthetic device, such as a PDA. We do not consider other natural body movements, such as facial expression or head movements, at this time. We limit ourselves to hand and arm movements.

In summary, then, gestures can be made by pointing or gesturing to objects and locations in the real world, or by interacting with a PDA display that represents the same environment. Furthermore, commands can be verbal or made by clicking on a PDA screen. In all cases, commands and gestures that the human user wishes to communicate to the robot are similarly translated into domain predicates. These predicates are stored and noted as either being completed or not. We are keeping a record of the actions or goals of the interactions so that we can address the issue of adjustable autonomy. We turn now to a discussion of the multi-modal system in greater detail, and then examine the issue of how adjustable autonomy is affected.

1.1 System overview

We have been designing and implementing a multi-modal interface to autonomous robots. In our research, we have been using Nomad 200s, XR-4000s, and an RWI ATRV-Jr. For a schematized overview of our system, see Figure 1.

We will limit our discussion here to the Nomad 200, its interactions with our in-house natural language processing system, Nautilus,[5.] and the handheld PDA, a Palm V Organizer. The robots understand speech, hand gestures, and input from a hand-held PDA.

Speech is initially processed by a speech-to-text voice recognition system (IBM's ViaVoice). Our natural language understanding system, Nautilus, robustly parses the resulting text string, translating it into a semantic representation which is then mapped to a robotic system command. During the processing, gestural information is incorporated.

Thus, for example, the acoustic signal for such a command as "Go to the waypoint over there" is translated to a string by our off-the-shelf speech recognition system. Nautilus parses the string and maps it to a semantic representation, something like (4).

(4)
```
(imper (:verb gesture-go
          (:agent (:system you))
          (:to-loc (waypoint-gesture))
          (:goal (there))) 0)
```

(1)

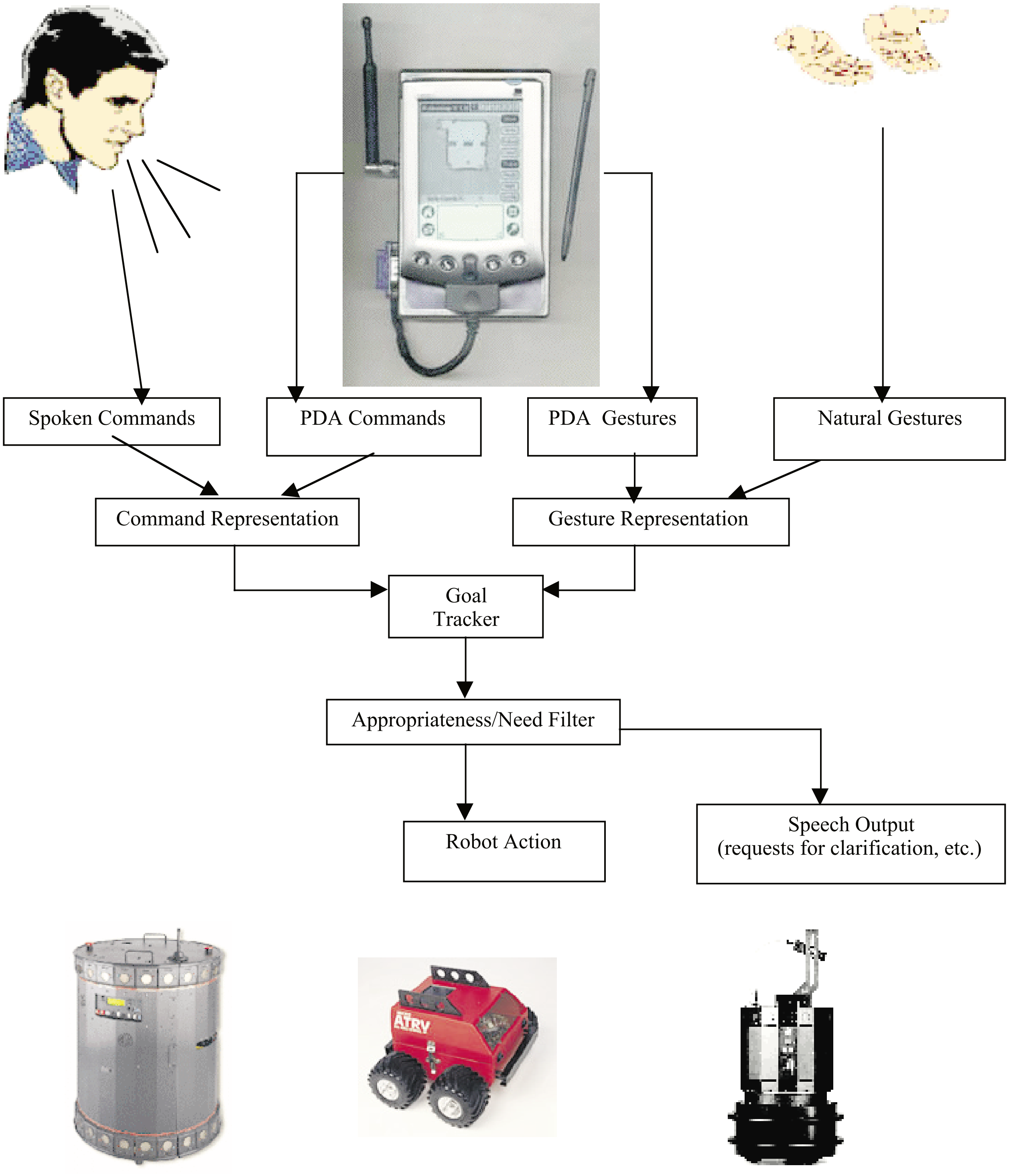

Figure 1: Schematic diagram of system architecture

(4) contains the main predicate or verb of the sentence and its arguments. In this case, the predicate is the verb "go" which is semantically classified in our system as a gesture-go type of verb. This verb semantically takes three arguments: an agent, a to-loc argument, and a goal. These three arguments are the mandatory subject of the command, "you," the object of the sentence, "waypoint," and the adverbial "there" respectively. For our purposes, we eliminate the word "over" which is overtly present in the actual command, simply because we feel it is redundant information in the adverbial expression, "over there." Furthermore, for the purposes of our translation functions, we have had to classify a waypoint semantically as either "waypoint" or "waypoint-gesture" simply to capture the difference between statements containing waypoints without gestures, as in "Go to waypoint one," vs. "Go to the waypoint over there." Since our knowledge base contains the information of a number of waypoints, we have had to add this additional distinction to allow the processing of such sentences.

The last item of the list, the digit 0, is placed in the representation to indicate that the goal has not yet been achieved. If, during the discourse, the goal of going to a particular waypoint is achieved, then the representation is updated. In this case, it is a simple matter of changing the 0 to a 1. While the processing is rather simple, it allows us to keep track of goals, achieved or unattained, so that later, fragmented input can be processed completely, or unattained goals can be looked up and processed with the purpose of completion in mind. Having this information on hand promotes the idea of adjustable autonomy, since the system is aware of what needs to be done, what has been done, and what yet needs to be accomplished. Users do not necessarily have to repeat information or remember what has transpired, since the system already keeps and tracks this information and can respond appropriately when needed.

The palm device, which dynamically presents an adaptive map of the robot’s environment, can also be used to give certain commands to the robots. Users can tap on menu buttons on the device's touch screen, or gesture (by tapping or dragging across a map of the environment on the PDA screen) to indicate places or areas for the robots. The map on the display comes directly from the robot via a mapping and localization module.[6.]

The semantic information of (4) and gestural information obtained from one of the sensors on the robot, the rangefinder or the PDA, are then combined and checked to see that a totally meaningful utterance has been obtained. Let us now consider how gestural information is incorporated.

Natural gestures can signify either distances, indicated by holding the hands apart to indicate a distance, or directions, indicated by tracing a line in the air. Natural gestures are detected using a structured light rangefinder which emits a horizontal plane of laser light 30 inches above the floor. A camera fitted with a filter tuned to the laser wavelength is mounted on its side. Given that the laser and camera mount are at a right angle, and the camera is tilted a fixed amount, the distance to a laser-illuminated point can be easily triangulated. Using this sensor, the robot is capable of tracking the user's hands and interpreting their motion as either vectors or measured distances. Similarly, synthetic gestures can signify either locations or paths, indicated by clicking on a particular set of x,y coordinates on the palm device's screen, or by dragging the PDA's stylus across the screen to indicate a path.

We believe our interface exhibits a fairly high level of integration as evidenced by how well the various modules share information in order to complete a task. For example, if the speech module needs information from the gesture module and can obtain that information readily, we would say that this system exhibits a greater degree of integration than one in which the modules do not have ready access to another module's information. The payoff for having information shareable is that the user can concentrate on communicating with the system, not with *how* to communicate with it. The system can get whatever information it needs, because it is available. Interacting with an autonomous vehicle, therefore, should be easy: simply interact with it naturally and let it do the work of putting all the pieces together.

We have been building a human-robot interface that allows for natural as well as mechanical interaction. We have constructed the interface so that it is responsible for integrating the information for the various input modalities. Users, therefore, are free to interact with it as they see fit.

2. DISCUSSION

2.1 Using a dialog-based interface to achieve greater autonomy

Autonomous robots, equipped with sufficient knowledge, should be able to go off on their own and complete actions without the human having to intervene at every step.

By tracking goals in the dialog during human-machine interactions, we are able to achieve a greater level of autonomy. Because the system knows whether or not a goal has been achieved, as in (4) above, by using an indicator in the representation for an utterance, the system can work on present actions, and if interrupted for some reason, can return to a previous action, simply by looking up the status of the previous goals and returning to any unattained ones. For example, suppose a robot is directed to go to a particular object or achieve some goal but is stopped for some reason before completing that goal. Next, suppose that the robot is then re-directed to achieve some other action, but during the completion of this action, unforeseen obstacles prevent the robot from completing this second goal. Clearly, some planning component is necessary to address the issue of handling the current problem. The issue of what to do upon its solution must also be addressed. We believe our multi-modal interface permits this activity.

Recent work in planning[7.] indicates that a planning component is necessary for collaborative work between multiple agents. Collaboration entails team members adjusting their autonomy through cooperation. While we will not discuss the intricacies of a planning module here, we have started to incorporate some elements of planning by attempting to use natural language dialog and goal tracking as a planning activity.

By generating plans from goals, and prioritizing them, almost on the fly as it were, the robotic system can achieve the kinds of coordination only obtainable by systems internally adjusting and cooperating with other systems that are themselves adapting to their role in a team and to a changing environment.[8., 9.] Once having addressed the immediate problem, namely solving the problem of confronting an unforeseen obstacle in the completion of a task, the robot is then free to check back to see where it was in the completion of the overall task and resume activity, based on the information it has about goals, as it has been tracking and updating them throughout the discourse. We will now consider a concrete example of how our system addresses this problem, specifically by considering one aspect of goal tracking in our dialog-based system. The result is a system that also exhibits greater autonomy.

2.1 An example

Clearly, we do not want to have systems that we must constantly spend our time tracking, keeping watch over, and constantly correcting when problems arise. We wish our systems to be adaptive and more independent and allow us freedom in interacting with them. We have addressed part of the problem of interaction with unmanned systems in one area above. We now wish to consider how a dialog-based system can further aid in achieving adjustable autonomy.

Consider the following scenario. A user is interacting with a robotic system. After initiating the goal, the human user leaves the scene to address another problem.

(5) User: "Go to the waypoint over there."

The user gestures to a particular location and then moves away to pursue some other independent activity. The robot, having received an appropriate command and accompanying gesture, proceeds to the destination. However, before the robot can obtain the desired location, the human user believes it has taken long enough for the robot to achieve the goal, and via an intercom system poses the following question.

(6) User: "Are you there yet?"

For the user, the locative reference "there" in (6) is to the location previously pointed to; namely "the waypoint over there" in the user's initial command (5). However, the robot, having various sensors, "sees" a different location at this point; namely, the location it is currently looking at. Therefore, the referent "there" in the user's query could easily be misconstrued as the location that the robot is currently looking at. This of course could lead to an erroneous positive response on the part of the robot, since it is at its present location.

However, by tracking the goals as we outlined above in (4), the system would be able to check back to see what exactly the user means by using the word "there." In such a tracking system along with a gesture recognition system as we have outlined here, the locative expression would clearly refer back to the coordinates which the user gestured to in the previous utterance. Next, the system could easily determine that the goal had not been achieved yet; therefore, an appropriate response would only be negative.

By keep tracking of goals in a dialog, our system has greater autonomy, since it is not necessary for user's to unnaturally repeat an argument, to intercede frequently to determine progress, correct certain kinds of misinterpretations which could have been avoided with a more robustly designed interpretation component, or to have to repeat commands to keep the system on track. By keeping tracking of goals on its own, our system is more autonomous and can complete its tasks whether interrupted or even if what might be perceived as confusing information is presented during the course of a dialog.

3. CONCLUSIONS

Interactive systems should firstly be easy to use. User should not have to worry about how to interact with the system; the system should take care of that. The user should be free to interact and use any mode of communication with the system however he/she sees fit. Secondly, interactive systems should be designed so that the robotic system is capable of adjusting its autonomy. Users should be allowed to interact with the system, believing that once a set of instructions is given, the system will monitor itself and only require minimal watching to ensure that goals are achieved.

Our work designing and implementing a multi-modal interface to autonomous robots is focussing on two broad research issues. Our research regarding the multi-modal interface is primarily concerned with the integration of the command and gesture modules of our interface. Furthermore, the interface should be so constructed as to allow the user complete freedom in determining how he/she wishes to interact with it. We believe considering integration as the impetus for constructing a multi-modal interface leads to a more natural and easier-to-use interface. On the robot side of the house, we are focussing on autonomy, trying to construct a system that knows enough about itself, the world around it, and what it has been doing, so that it becomes more of a team player when interacting with humans and other robots. We are focussing our research on building a dialog-driven planner to track goals during human-robot interactions. When necessary, robots should be totally autonomous and carry out their duties and functions; however, they should be adaptable to any situation as it arises, becoming more dependent if necessary, acting closely with their team members as situations may demand.

ACKNOWLEDGMENTS

This work is funded in part by the Naval Research Laboratory and the Office of Naval Research.

REFERENCES

1. B. Grosz, and C. Sidner, "Attention, Intentions, and the Structure of Discourse" *Computational Linguistics* 12(3), pp. 175-204, 1986.
2. D. Perzanowski, A.C. Schultz, and W. Adams, "Integrating Natural Language and Gesture in a Robotics Domain.," *Proceedings of the IEEE International Symposium on Intelligent Control*, pp. 247-252. National Institute of Standards and Technology, Gaithersburg, MD, 1998.
3. D. Perzanowski,, A. Schultz, E. Marsh, and W. Adams, "Goal Tracking in a Natural Language Interface: Towards Achieving Adjustable Autonomy," *Proceedings of the 1999 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp. 144-149. IEEE Press, Monterey, CA, 1999.
4. D. Perzanowski, A. Schulz, W. Adams, and E. Marsh, "Towards Seamless Integration in a Multi-modal Interface," *Proceedings of the 2000 Workshop on Interactive Robotics and Entertainment*, AAAI Press, Pittsburgh, PA, to appear.
5. K. Wauchope, "Eucalyptus: Integrating Natural Language Input with a Graphical User Interface," Technical Report, NRL/FR/5510--94-9711, Washington, DC, Naval Research Laboratory, 1994.
6. A. Schultz, W. Adams, and B. Yamauchi, "Integrating Exploration, Localization, Navigation and Planning With a Common Representation," *Autonomous Robots,* **6(3***),* pp. 293-308, 1999.
7. B. Grosz, L. Hunsberger, and S. Kraus, "Planning and Acting Together," *AI Magazine* **20(4),** pp. 23-34, 1999.
8. M. Pollack, and J.F. Horty, "There's More to Life Than Making Plans," *AI Magazine* **20(4)**, pp. 71-83, 1999.
9. M. Pollack, and C. McCarthy, "Towards Focused Plan Monitoring: A Technique and an Application to Mobile Robots," *Proceedings of the 1999 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp. 144-149, IEEE Press, Monterey, CA, 1999.